Федеральное государственное автономное образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ:
заведующий кафедрой

_____ В. А. Кратасюк
“___” _____ 2020 г.

БАКАЛАВРСКАЯ РАБОТА

03.03.02 Физика

ВЫЯВЛЕНИЕ СВЯЗЕЙ МЕЖДУ ТАКСОНОМИЕЙ, ФУНКЦИЕЙ И ТРИПЛЕТНЫМ СОСТАВОМ МИТОХОНДРИАЛЬНЫХ ГЕНОВ НЕКОТОРЫХ ГРИБОВ

Руководитель: _____ д.ф.-м.н., проф. М. Г. Садовский
дата, подпись; уч.степень, должность

Выпускник: _____ В. Д. Федотовская
дата, подпись

Красноярск 2020

РЕФЕРАТ

Выпускная квалификационная работа по теме «Выявление связей между таксономией, функцией и триплетным составом митохондриальных генов некоторых грибов» содержит 45 страниц текстового документа, 30 использованных источников, 4 рисунка, 6 таблиц.

СТРУКТУРА, ФУНКЦИЯ, НУКЛЕОТИДНЫЕ ПОСЛЕДОВАТЕЛЬНОСТИ, УПРУГИЕ КАРТЫ, МЕТОД ДИНАМИЧЕСКИХ ЯДЕР, МИТОХОНДРИАЛЬНЫЕ ГЕНОМЫ ГРИБОВ

Цель работы — анализ распределения генов митохондрий грибов в пространстве частот триплетов.

Выявление связи между структурой нуклеотидной последовательности и кодируемой ей функцией является одной из основных задач биоинформатики. Эта связь изучалась на примере трех белок-кодирующих генов митохондрий грибов: *atp*6, *atp*8, *atp*9. Из полных митохондриальных геномов извлекались последовательности указанных генов в двух версиях: зрелые мРНК и исходные последовательности с интронами. Каждая группа генов преобразовывалась в частотный словарь триплетов в двух вариантах: с шагом рамки считывания $t = 1$ и $t = 3$. Далее проводилась кластеризация словарей в пространстве частот триплетов линейным (метод динамических ядер) и нелинейным (метод упругих карт) методами. Результат кластеризации проверялся на устойчивость.

В данной работе рассматривались два ключевых вопроса: верно ли, что словари образуют кластеры в пространстве частот триплетов? И если да, то кластеры содержат словари, соответствующие одинаковым генам или близким таксонам? В результате обнаружилось, что при кластеризации всех типов словарей, образуются функционально специфичные кластеры. Таким образом было доказано преобладание функции, кодируемой последовательностями, над таксономией ее носителей.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Изучение любого генетического материала (генома митохондрий в нашем случае) требует его анализа с трёх сторон: анализ его структуры, его функции и таксономии его носителя. Каждую из этих сторон можно изучать индивидуально, однако большой интерес представляет их сочетанный анализ. Именно такому анализу посвящена настоящая работа.

Актуальность работы связана с появлением большого количества данных, а также с быстро развивающимися технологиями секвенирования генетического материала, что дает возможность быстро и эффективно изучить отличительные характеристики организмов. Тем не менее, геномы митохондрий грибов изучены хуже, чем геномы митохондрий растений. Митохондриальная ДНК может быть полезна в изучении эволюции грибов, а также в проведении популяционного анализа и построении филогении. Кроме того, некоторые грибы являются патогенами, вызывающими различные болезни лесных сообществ. Они могут нанести значительный, а иногда и непоправимый вред древесным растениям. Изучение патогенных организмов возможно с нескольких сторон, в том числе и с точки зрения генетики.

Объектом настоящей работы является связь между структурой, функцией и таксономией носителя определённых генов митохондрий грибов.

Предметом настоящей работы являются:

- последовательности генов митохондрий грибов 223 видов пяти таксономических отделов: *Ascomycetes* (185 видов), *Basidiomycetes* (24 вида), *Blastocladiomycota* (2 вида), *Chytridiomycota* (6 видов) и *Zygomycota* (6 видов);
- структура этих генов, определяемая их триплетным составом;
- связь между выделяемыми неоднородностями в этом распределении, таксономией и функцией.

Целью данной работы является выявление связи между триплетным составом нуклеотидной последовательности, ее функцией и таксономией ее носителей на примере генов митохондрий грибов. Для выявления такой связи были выбраны следующие гены: *atp*6, *atp*8 и *atp*9. Для достижения данной цели были поставлены следующие **задачи**:

1) Создание базы нуклеотидных последовательностей;
2) Построение частотных словарей последовательностей;
3) Кластеризация словарей методом упругих карт;
4) Кластеризация словарей методом динамических ядер ($2 \leqslant K \leqslant 5$);
5) Анализ распределения словарей по кластерам с точки зрения функционального и таксономического состава.

Работа докладывалась на следующих конференциях:

- 56-я Международная научная студенческая конференция, Новосибирск, устный доклад;
- X международная конференция «Dynamical Systems Applied to Biology and Natural Sciences» (DSABNS), Неаполь, стендовый доклад;
- Международная конференция студентов, аспирантов и молодых ученых «Проспект Свободный — 2019», Красноярск, устный доклад;
- 7^{th} International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO), Гранада, устный доклад;
- XI международная конференция «Dynamical Systems Applied to Biology and Natural Sciences» (DSABNS), Тренто, стендовый доклад;
- Международная конференция студентов, аспирантов и молодых ученых «Проспект Свободный — 2020», Красноярск, устный доклад.

Результаты работы опубликованы в следующих научных журналах и сборниках научных мероприятий:

- Колесникова А.И. Выявление связи тринуклеотидного состава генов и таксономии их носителей на примере генов митохондрий некоторых грибов / Колесникова А.И., Федотовская В.Д., Шпагина Т.О. // Материалы 56-й Международной научной студенческой конференции (МНСК). — 2018. — Vol. 56. — Стр. 18;

- Колесникова А.И. Влияние функциональных различий сильнее влияния таксономических различий для генов семейства atp митохондрий грибов / Колесникова А.И., Федотовская В.Д., Шпагина Т.О., Садовский М.Г. // Моделирование неравновесных систем. / Материалы XXI Всероссийского семинара — 2018. — Vol. 21. — Стр. 49–54;
- Федотовская В. Д. О соотношении влияния функциональных и таксономических различий для генов семейства atp митохондрий некоторых грибов / Федотовская В.Д., Шпагина Т.О., Колесникова А.И., Садовский М.Г. // Нейроинформатика, ее приложения и анализ данных. / Материалы XXVII Всероссийского семинара — 2019. — Vol. 27. — Стр. 110–115;
- Sadovsky M., Fedotovskaya V., Kolesnikova A., Shpagina T., Putintseva Y. Function vs. Taxonomy: The Case of Fungi Mitochondria ATP Synthase Genes // Lecture Notes in Computer Science, Springer Verlag. — 2019. — Vol. 11465 — Pp. 335–345;
- Fedotovskaya V. Kolesnikova A., Shpagina T., Putintseva Y., Sadovsky M. Function Overcomes Taxonomy: Case of ATP Genes of Fungi Mitochondria // Tenth International Conference Dynamical Systems Applied to Biology and Natural Sciences: Book of Abstracts. — 2019. — Vol. 10. — Pp. 176–176.
- Fedotovskaya V. Kolesnikova A., Shpagina T., Sadovsky M. The Distribution of Fungal Mitochondrial ATP Genes in Amino Acids Space // 11^{th} International Conference Dynaamical Systems Applied to Biology and Natural Sciences: Book of Abstracts. — 2020. — Vol. 11. — Pp. 215-216.
- Fedotovskaya V., Sadovsky M., Kolesnikova A., Shpagina T., Putintseva Y. Function vs. taxonomy: further reading from fungal mitochondrial xtitATP synthases // Lecture Notes in Computer Science, Springer Verlag. — 2020. — Vol. 12108 — Pp. 438–444.

1 Обзор литературы

1.1 Митохондрии

Митохондрии — палочкообразные клеточные органеллы, диаметром около 1 мкм и длинной до 7 мкм, встречающиеся только в эукариотических клетках. Митохондрии имеют двойную мембрану. Область, ограниченная складчатой внутренней мембраной (складки называются кристами) и известная под названием митохондриального матрикса, содержит рибосомы и митохондриальную ДНК — кольцевую двухцепочечную молекулу, кодирующую некоторые митохондриальные белки. Во внутренней мембране локализован фермент, ответственный за синтез АТФ, — так называемый F1-комплекс. В компартменте, заключенном между наружной и внутренней мембранами, находятся субстраты, ферменты и некоторые метаболиты. Число митохондрий в одной клетке может достигать нескольких тысяч. Их основной функцией является окисление органических соединений и синтез АТФ с использованием энергии окисляющихся веществ. Происходит это благодаря кристам митохондрий, которые участвуют в системе переноса электронов. Молекула АТФ является наиболее универсальным источником энергии, хотя в некоторых реакциях могут использоваться и другие нуклеозидтрифосфаты, например, ГТФ. АТФ образуется в цитоплазме в результате разнообразных катаболических процессов, таких как гликолиз, а энергия, запасенная в трифосфатной группе, расходуется в ходе биосинтетических реакций.

Прокариотический характер генетических систем митохондрий (а также хлоропластов) указывает на то, что эта органелла эволюционировала из бактерий, которые эндоцитировали более 1 миллиарда лет назад. Согласно одной из версий этой эндосимбиотической гипотезы, эукариотические клетки изначально были анаэробными организмами без митохондрий, но затем вступили в стабильные эндосимбиотические отношения с бактерией, систему окислительного фосфорилирования которой они приспособили к собственным нуждам. Анализ сравнения последовательностей показывает, что митохондрии произошли от определенного типа пурпурных фотосин-

тетических бактерий, которые ранее потеряли способность фотосинтезировать, и у которых осталась только дыхательная цепь [1].

1.2 Геномы митохондрий

Чаще всего митохондриальная ДНК представлена в виде кольцевой двухцепочечной молекулы. Реже встречаются линейные ДНК, например, у рода паразитических одноклеточных организмов *Plasmodium*. Многочисленные копии ДНК находятся в матриксе митохондрий и обычно распределены по нескольким кластерам, называемым нуклеоидами. Считается, что нуклеоиды прикреплены к внутренней мембране. Пространственная структура ДНК в нуклеоидах скорее напоминает структуру ДНК бактерий, чем эукариотический хроматин; например, как и у бактерий, в органеллах отсутствуют гистоны. Молекулы митохондриальной ДНК меняются в размерах от менее 600 пар нуклеотидов в *Plasmodium falciparum* (малярийный паразит человека) до более 10^6 п. н. в некоторых наземных растениях [2,3].

Как митохондриальные, так и ядерные геномы многих эукариот помимо генов содержат избыток разбросанных по геному некодирующих последовательностей. Кодирующие участки генома в составе гена называются экзонами. Некодирующие участки, расположенные между экзонами, называются интронами. Кроме того, между генами также находятся некодирующие участки генома. Такие участки генома называют иногда «мусорной» ДНК, так как их полезность для клетки не была доказана. Тем не менее, часть этой ДНК может участвовать в экспрессии некоторых генов, играть роль регуляторных сигналов. Сначала с каждого гена синтезируется пре-РНК, в которой содержатся как интроны, так и экзоны. После чего интроны удаляются, а экзоны сшиваются в одну цепь зрелой иРНК. Этот процесс называется сплайсингом. В процессе альтернативного сплайсинга некоторые участки интронов могут сохраняться в зрелой мРНК.

Некоторые митохондриальные геномы растений и грибов содержат интроны, что является неожиданным, поскольку они нечасто встречаются в генах бактерий, от чьих предков, как предполагается, произошли мито-

хондрии. В дрожжах один и тот же митохондриальный ген может содержать интрон в одном штамме, но не содержать в другом. Такие интроны, по-видимому, способны покидать геном и возвращаться в него как мобильные генетические элементы [1].

Митохондриальная ДНК грибов находится в матриксе митохондрий. Чаще всего она представлена кольцевой двухцепочечной молекулой, кодирующей 14 основных генов: *atp*6, *atp*8, *atp*9, *cob*, *cox*1, *cox*2, *cox*3, *nad*1, *nad*2, *nad*3, *nad*4, *nad*5, *nad*6 и *nad*4L, отвечающих за окислительное фосфорилирование. Из них:

- гены *nad*1, *nad*2, *nad*3, *nad*4, *nad*4L, *nad*5 и *nad*6 отвечают за транспорт электронов I комплекса;
- гены *cox*1, *cox*2, *cox*3 отвечают за транспорт электронов IV комплекса;
- ген *cob* отвечает за перенос электронов III комплекса;
- гены комплекса АТФ-синтазы *atp*6, *atp*8, *atp*9.

В данной работе изучаются гены *atp*6, *atp*8 и *atp*9, которые отвечают за окислительное фосфорилирование и синтез АТФ.

1.3 Окислительное фосфорилирование

В эукариотических клетках окислительное фосфорилирование протекает в митохондриях. Митохондрии имеют две мембраны. Внешняя мембрана хорошо проницаема для низкомолекулярных соединений, а внутренняя непроницаема для многих соединений, лишь некоторые могут пройти через внутреннюю мембрану с помощью специальных переносчиков. Благодаря такой избирательной проницаемости внутренняя мембрана разграничивает метаболические процессы, протекающие в цитозоле от протекающих в матриксе.

Первый этап окислительного фосфорилирования — поступление электронов в дыхательную цепь. Электроны доставляются дегидрогеназами к универсальным акцепторам электронов — никотинамиднуклеотидам $НАД^+$ или $НАДФ^+$ и флавиннуклеотидам (ФМН или ФАД).

Дыхательная цепь переноса электронов состоит из ряда последова-

тельно действующих переносчиков электронов. Перенос электронов может осуществляться тремя способами:

- Путем прямого переноса;
- Путем переноса атома водорода;
- Путем переноса гидрид-иона.

Кроме никотинамиднуклеотидов и флавиннуклеотидов в переносе электронов в дыхательной цепи также участвуют убихинон и железосодержащие белки (цитохромы и железо-серные белки). Процесс переноса электронов коротко можно описать следующим образом. Электроны от первичных доноров попадают на флавопротеины, убихинон и далее по цепи железо-серных белков и цитохромов в конце концов попадают на молекулярный кислород.

Из-за переноса электронов возникает градиент H^+. Такой электрохимический градиент называется протон-движущей силой, которая служит источником энергии для синтеза АТФ из АДФ.

1.4 Методы выделения структурированности данных

Изучить большой объем данных и выделить его структурированность позволяют методы кластерного анализа. Это статистические методы, выполняющие разделение множества входных данных на группы на основе их сходства друг с другом. К настоящему времени существует огромное число различных алгоритмов кластерного анализа, однако общепринятой классификации пока нет. Тем не менее, можно выделить следующие группы подходов. Часть методов может быть отнесена сразу к нескольким группам, поэтому приведенное ниже разделение можно рассматривать лишь как приближение к реальной классификации.

Первая группа — иерархические алгоритмы. Такие алгоритмы основаны на создании иерархии или представлении вложенных кластеров в виде графа или дендрограммы. Построение графов происходит в том предположении, что множество изучаемых точек, объединяемых в отдельный кластер, характеризуется их «близостью». Выделяют две разновидности

иерархических алгоритмов: агломеративные и дивизивные методы. Агломеративные методы предполагают объединение более мелких кластеров в более крупные: дерево строится от листьев к стволу. В начале анализа массива данных все точки являются отдельными классами. На первом шаге наиболее похожие элементы объединяются в отдельные кластеры. На последующих шагах мелкие кластеры объединяются в крупные до тех пор, пока все точки не будут принадлежать одному кластеру. Однако процесс должен быть остановлен, когда последующее объединение будет приводить к нежелательным кластерам. Например, можно остановиться, когда точки, объединенные в новый кластер, разбросаны по достаточно большой области. Дивизивные методы, наоборот, предполагают выделение более мелких кластеров из более крупных. Таким образом, дерево строится от ствола к листьям. Примерами иерархических алгоритмов могут служить методы одиночной связи, полной связи и средней связи и другие. Различие указанных методов заключается в выборе определения расстояния между кластерами.

Эти методы имеют существенный недостаток: результат классификации чувствителен к изъятию точек. Существуют такие конфигурации данных, что при изъятии даже одной точки, результат классификации будет сильно отличаться от полученного первоначально.

Рост объёма анализируемых данных делает иерархические алгоритмы кластеризации непригодными. Поэтому выделяется вторая группа методов — статистические алгоритмы, главная идея которых заключается в том, что кластер хорошо описывается некоторым вероятностным распределением. Эти методы направлены на оптимизацию какой-либо выбранной исследователем функции, позволяющей выделить оптимальное распределение данных на кластеры. В качестве примеров статистических алгоритмов можно указать метод динамических ядер, EM-алгоритм и другие.

Стоит отметить, что единого универсального алгоритма кластеризации не существует: каждый имеет свои достоинства и недостатки, которые нужно учитывать. Алгоритм кластеризации выбирается исходя из

Таблица 1 — Некоторые функции расстояния

Название	Определение
Линейное расстояние	$d(I^{(i)}, I^{(j)}) = \sum_{k=1}^{m} \left\lvert I_k^{(i)} - I_k^{(j)} \right\rvert$
Евклидово расстояние	$d(I^{(i)}, I^{(j)}) = \sqrt{\sum_{k=1}^{m} \left(I_k^{(i)} - I_k^{(j)} \right)^2}$
Квадрат евклидова расстояния	$d(I^{(i)}, I^{(j)}) = \sum_{k=1}^{m} \left(I_k^{(i)} - I_k^{(j)} \right)^2$
Обобщенное степенное расстояние Минковского	$d(I^{(i)}, I^{(j)}) = \left[\sum_{k=1}^{m} \left(I_k^{(i)} - I_k^{(j)} \right)^p \right]^{\frac{1}{p}}$

поставленных задач, природы данных, а также исходя из планирующихся результатов. В зависимости от выбранного алгоритма может выбираться разное количество кластеров, а также функция расстояния между объектами. Окончательный результат кластеризации во многом зависит от выбора функции расстояния. В таблице 1 приведены некоторые из них [4].

В данной работе методы кластерного анализа использовались для выделения групп генов, которые склонны образовывать скопления, а также для выявления признака, по которому частотные словари попадают в один и тот же кластер: по одинаковым генам или по близким таксонам. Для выполнения поставленных задач были выбраны два метода кластеризации: линейный и нелинейный. Для линейной кластеризации использовался метод динамических ядер, а для нелинейной — метод упругих карт. Описание данных методов приведено в Главе 2.

1.5 Выбор генетического материала и анализируемых структур

Данное исследование существенным образом зависит от того, какой именно генетический материал берётся в рассмотрение. В работах [5,6] исследовалась связь между таксономией носителя и геномов органелл: хлоропластов и митохондрий, соответственно. Такой выбор генетического материала позволяет редуцировать задачу: функция у всех хлоропластов (митохондрий, соответственно) одна и та же. В работах [7,8] в качестве генетического материала использовались последовательности зрелых РНК генов 16 S РНК бактерий; такой выбор генетического материала так же позволял

рассматривать лишь связь структура — таксономия.

Изучению структуры и свойств биологических объектов посвящено большое количество работ. Так, например, в работе [9] было проведено сравнение таксономической структуры кормовых растений западного Забайкалья. В работах [7, 8, 10] изучалась классификация нуклеотидных последовательностей бактериальных РНК.

Также разнообразие структур, выделяемых исследователями в последовательностях ДНК, весьма велико. Например, в работе [11] в качестве структуры рассматривался частотный словарь триплетов, кластеризация проводилась методом динамических ядер. В работах [12, 13] также использовались частотные словари, но визуализация и последующий анализ проводились с помощью метода главных компонент.

2 Материалы и методы

2.1 Генетический материал

Генетический материал был взят из открытой базы данных, содержащей все аннотированные на сегодняшний день последовательности ДНК, РНК и белков, NCBI GenBank[1]. Далее, с помощью программы CLC Genomics Workbench из исходной последовательности были выделены выбранные для исследования гены в двух вариантах: зрелые мРНК — кодирующие последовательности, и последовательности, содержащие как экзоны, так и интроны. Впредь кодирующие последовательности будут называться CDS, а последовательности, содержащие как кодирующие, так и некодирующие участки — Gene. Выбор данных вариантов последовательностей объясняется тем, что последовательности, содержащие кодирующие и некодирующие участки, находятся в геноме именно в таком виде, а последовательности, состоящие только из экзонов, кодируют зрелую РНК, готовую к трансляции. В Таблице 2 представлено количественное соотношение исследуемой выборки организмов по отделам.

Таблица 2 — Состав выборки организмов по отделам; N — число организмов.

Отдел	N
Ascomycetes	185
Basidiomycetes	24
Blasticladiomycota	2
Chytridiomycota	6
Zygomycota	6

2.2 Частотные словари

Целью данной работы является выявление связей и закономерностей в триаде *структура – функция – таксономия*. При изучении таких связей исследователь, как правило, однозначно понимает, что является функцией, определяемой последовательностью, и что является таксономическим положением носителя данной последовательности. Однако, структура нуклеотидной последовательности может истолковываться по-разному. В настоящей работе под структурой будем понимать триплетный состав генов,

[1] https://www.ncbi.nlm.nih.gov

а именно, различные частотные словари, которые будут описаны ниже.

Любая нуклеотидная последовательность — это символьная последовательность из алфавита, содержащего 4 буквы: $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. Все исследуемые последовательности были разделены на подпоследовательности длиной q. Сами подпоследовательности (всевозможные триплеты от $\omega_1 = \mathsf{AAA}$ до $\omega_{64} = \mathsf{TTT}$) и информацию о частоте их встречаемости в исходной последовательности назовем частотным словарем W_q. Частотные словари были построены программой *ad hoc*, написанной на языке программирования *Python*.

Частотные словари были составлены в двух версиях. Опишем кратко общую схему составления. Рамку считывания длиной $q = 3$ поместим в начало нуклеотидной последовательности и будем перемещать ее вдоль этой последовательности с шагом $t \geqslant 1$. При этом будем вести список всех встретившихся триплетов ω, а также подсчитывать количество копий n_ω. Частота конкретного триплета может быть рассчитана по формуле:

$$f_\omega = \frac{n_\omega}{M}\,, \tag{1}$$

где M — количество всех встретившихся триплетов. Совокупность частот всех триплетов есть частотный словарь $W_{(q,t)}$. Далее в работе изучались частотные словари $W_{(3,1)}$ и $W_{(3,3)}$. Следует отметить, что такая структура неоднозначно соответствует гену: у разных генов могут быть совпадающие частотные словари, при этом однозначный обратный переход от частотного словаря в последовательность гена невозможен.

Частотные словари являются точками в многомерном пространстве размерности равной 64. В этом пространстве можно задать различные метрики. В данной работе использовалась Евклидова метрика, расстояние в которой определяется следующим образом:

$$d\left(W^{(1)}, W^{(2)}\right) = \sqrt{\sum_{\omega=\mathsf{AAA}}^{\mathsf{TTT}} \left(f_{\omega^{(1)}} - f_{\omega^{(2)}}\right)^2} \tag{2}$$

Здесь $W^{(1)}$ и $W^{(2)}$ — частотные словари, $f_\omega^{(1)}$ и $f_\omega^{(2)}$ — частоты триплета ω в первом и во втором частотном словаре, соответственно.

2.3 Метод динамических ядер

Метод динамических ядер, известный в англоязычной литературе как K–means, является одним из наиболее известных методов кластеризации без учителя. Это итерационный метод, основанный на переопределении центра масс, полученного на предыдущем шаге. Прежде всего нужно выбрать число классов, на которые следует поделить множество исследуемых точек; пусть оно будет равно K. На первом шаге все точки делятся случайно по K классам.

Для каждого класса определяется его динамическое ядро — среднее арифметическое всех точек, попавших в этот класс. После для каждого динамического ядра вычислим расстояние для каждой точки и переопределим принадлежность точек по правилам:

1. Если точка первоначально принадлежала i–ому классу, но оказалась ближе к центу n–ого класса, то переместим ее в n–ый класс;
2. В противном случае оставим эту точку в первоначальном классе.

После того, как все точки были переопределены, заново вычисляются динамические ядра классов, и повторяется процедура переопределения принадлежность точек к классам. Эта процедура повторяется до тех пор, пока принадлежность точек к тому или иному классу не остается неизменной. Стоит отметить, что эта процедура не всегда останавливается за конечное число шагов, однако такая конфигурация точек, при которой метод динамических ядер не останавливается, не является структурно устойчивой.

Строго говоря, после получения классификации стоит проверить различимость классов. Для такой проверки существуют два критерия различимости: сильный и слабый. Сильный требует, чтобы для различимости любых двух классов сумма их радиусов была не меньше расстояния между их центрами. Слабый требует, чтобы радиус большего класса был не больше расстояния между классами. Если два каких-то класса не прошли проверку различимости, их объединяют в один и всю процедуру повторяют заново. Таким образом, метод динамических ядер не увеличивает число

классов и останавливается на наибольшем количестве различимых классов.

Как правило, метод динамических ядер не дает один и тот же результат классификации на одних и тех же данных в двух разных реализациях. Причина этого состоит в том, что при каждой реализации метода на одних и тех же данных, начальное разделение на классы случайно и не совпадает друг с другом. В связи с этим возникает вопрос об устойчивой классификации. Так как классификация начинается со случайного распределения данных по классам, результат классификации будет зависеть от этого распределения. Если же разделение точек по классам всегда принимает один и тот же вид независимо от начального распределения точек, то такая классификация является устойчивой, но в реальных данных встречается это крайне редко.

Для того, чтобы ответить на вопрос об устойчивости классификации, следует провести достаточно длинную серию повторных реализаций метода динамических ядер на одних и тех же данных, но с различным начальным распределением точек по классам. После чего может возникнуть несколько вариантов исходов:

- во-первых, каждый запуск метода может привести к одному и тому же результату классификации. Такой вариант соответствует устойчивой классификации, обсуждаемой выше;
- во-вторых, каждый запуск метода может привести к абсолютно разным результатам. Такой случай считается полностью неустойчивым;
- в-третьих, большая часть результатов классификации приводит к одной и той же конфигурации точек в классах, тогда как остальная часть результатов может быть либо совершенно разными, либо одинаковыми. При этом, может быть такое, что большая часть точек, как правило, образует одну и ту же конфигурацию, вместе с тем остальная часть довольно случайно меняет свою принадлежность к классу. Этот случай стоит считать устойчивым, но необходимо ввести какой-либо критерий устойчивости, например, долю реализаций, приводящей к одному и тому же результату кластеризации.

Изучение точек, меняющих свою принадлежность к тому или иному классу при проведении кластеризации методом динамических ядер, может иметь важное значение для исследования. В таком неустойчивом поведении точек могут наблюдаться свои закономерности, которые, возможно, будут иметь свое биологическое значение.

2.4 Метод упругих карт

В данной работе методы кластерного анализа использовались для выделения групп генов, которые склонны образовывать скопления, определяемые близостью частот одинаковых триплетов. Ключевой задачей является проверка того, попадают ли в один кластер словари, соответствующие одинаковым (близким) генам, либо близким таксонам.

Кроме метода динамических ядер для выполнения поставленных задач был выбран нелинейный метод кластеризации — метод упругих карт. Упругая карта служит для сокращения размерности данных и является современным методом нелинейного статистического анализа. По своей математической сути, этот метод является методом замены переменных. В многомерном пространстве находится по возможности не слишком изогнутая поверхность, на которую проецируются данные, в результате чего они могут отображаться на ней как на карте. Эту поверхность удобно представить, как упругую пластину, находящуюся в пространстве данных, прикрепленных к ней пружинами. Метод упругих карт является обобщением метода главных компонент, в котором вместо упругой пластины используется абсолютно жесткая плоскость [14–16].

Пусть имеется множество точек в многомерном метрическом пространстве. На первом шаге определяются первая и вторая главные компоненты. Первая главная компонента — это такое направление в пространстве, вдоль которого наблюдается максимальный разброс изучаемых данных. Вторая главная компонента выбирается перпендикулярно первой в следующем по величине направлении максимального разброса.

Далее на определенных главных компонентах как на осях строится

плоскость, на которую проектируются изучаемые данные. Затем каждая точка соединяется со своей проекцией математической пружиной, обладающей бесконечной растяжимостью, причем ее свойства не меняются по мере растяжения. Следующим шагом определяется минимальный квадрат, охватывающий все проекции точек, стороны которого параллельны главным компонентам. Жесткая поверхность квадрата заменяется мембраной, которая обладает гибкостью и растяжимостью, а также может деформироваться. Необходимо отметить, что такая мембрана не должна разрываться и склеиваться.

Далее вся система отпускается, и в результате она приходит в состояние с минимальной энергией деформации. После этого происходит переопределение положения точек на мембране. Для этого определяются ортогональные проекции для каждой точки на деформированной поверхности. На последнем этапе деформированная эластичная поверхность освобождается от пружин и релаксирует до плоского состояния. В итоге эта поверхность снова становится квадратом. То, что получилось в результате описанных действий, называется представлением упругой карты во внутренних координатах [14–16].

Методы кластеризации, использованные в данной работе, основаны на определении локальной плотности точек, которая определяется как среднее число точек на небольшом участке упругой карты[2]. Для этого каждой точке сопоставляется какая-либо куполообразная функция распределения с максимумом в этой точке, например, функция Гаусса [17]

$$f_j(r) = A \cdot \exp\left\{-\frac{(r - r_j)^2}{\sigma^2}\right\} . \tag{3}$$

Здесь r_j — координата j–ой точки, A — множитель, одинаковый для всех точек, σ — полуширина этой функции, полностью аналогичная стандартному отклонению для случая нормального распределения случайной величины, подгоночный параметр, определяющий контрастность картины локальной плотности. Далее вычисляется функция, которая является суммой

[2]Во внутренних координатах.

всех определенных функций

$$F(r) = \sum_{j=1}^{N} \exp\left\{-\frac{(r - r_j)^2}{\sigma^2}\right\} . \tag{4}$$

Здесь N — количество точек. Для кластеризации и визуализации словарей методом упругих карт использовалось свободно распространяемое ПО *ViDaExpert*[3] [14–16].

[3] http://bioinfo-out.curie.fr/vidaexpert/

3 Результаты

3.1 Кластеризация методом динамических ядер

Множество точек каждого варианта словарей было последовательно разделено на 2, 3, 4 и 5 классов. Результаты такого деления для всех четырех типов словарей и изучаемого генетического материала представлены в Таблицах 3 — 6. Номера классов, приведенных в данных классификаци-

Таблица 3 — Распределение генов по двум классам в разрезе отделов. G — количество генов, попавших в класс; St — индекс устойчивости.

Отдел	G	St	G	St	G	St	G	St
	Класс I		Класс II		Класс I		Класс II	
	$W_{(3,1)}$, CDS				$W_{(3,1)}$, gene			
Ascomycetes	155	0,98	164	1	164	0,99	151	0,99
Basidiomycetes	24	0,98	24	1	24	0,99	24	0,99
Blasticladiomycota	2	0,98	2	1	2	0,99	2	0,99
Chytridiomycota	4	0,98	4	1	4	0,99	4	0,99
Zygomycota	6	0,98	6	1	6	0,99	6	0,99
	$W_{(3,3)}$, CDS				$W_{(3,3)}$, gene			
Ascomycetes	155	0,97	164	1	30	0,94	168	0,90
Basidiomycetes	24	0,97	24	1	5	0,94	24	0,90
Blasticladiomycota	2	0,97	2	1	0	0,94	2	0,90
Chytridiomycota	4	0,97	4	1	2	0,94	4	0,90
Zygomycota	6	0,97	6	1	3	0,94	6	0,90

ях, присваивался случайно, то есть важно не то, в какой по номеру класс попадает та или иная точка, а с какими из остальных точек она образует класс.

Как уже было сказано выше, ряд последовательных реализаций метода динамических ядер, как правило, не будет приводить к одному и тому же результату распределения на классы, в связи с чем возникает проблема определения устойчивости кластеризации. Эта проблема может решаться следующим способом: следует изучать такую конфигурацию разделения, которая является наиболее устойчивой, то есть когда часть точек часто

(например, не менее, чем в 70 % всех реализаций метода) попадает в один и тот же кластер, а другая часть постоянно меняет свое положение. Для того, чтобы проверить устойчивость классификации в данной работе, проводилось 100 последовательных запусков метода, после чего рассчитывался индекс устойчивости для каждого из выделенных классов. Индекс устойчивости St — доля тех реализаций метода динамических ядер, в которых наблюдалась одна и та же финальная конфигурация точек.

Кластеризация методом динамических ядер проводилась на трех генах АТФ-синтаз всех организмов. В Таблицах 3 – 6 представлены те гены, которые с определенной частотой (индекс устойчивости St) попадали в один и тот же класс. Остальные же гены не представлены в указанных таблицах, так как они часто меняли свою принадлежность к тому или иному классу. Такие гены объекты были названы волатильными. Важно, что в нашем случае их доля достаточно мала.

3.2 Иерархический характер кластеризации словарей методом динамических ядер

Результаты, показанные в Таблицах 3 – 6, не дают исчерпывающей характеристики связи между структурой, функцией и таксономией. В связи с этим был просмотрен состав выделенных кластеров с точки зрения находящихся в кластере генов. Для этого была введена наглядная конструкция, называемая слоистым графом. Графом является математический объект, состоящий из точек, называемых вершинами, и отрезков (ребер), соединяющих эти вершины. Слоистый граф — это такой граф, в котором множество всех вершин поделено на непересекающиеся подмножества, называемые слоями, при этом ребра не соединяют вершины внутри слоя или через слой, а только вершины соседних слоев. Состав кластеров при делении методом динамических ядер представлен на Рисунках 1–2.

В представленном случае вершинами являются кластеры, выделяемые методом динамических ядер; первый слой этих графов является результатом деления точек на два кластера, второй слой — результат деления

Таблица 4 — Распределение генов по трем классам в разрезе отделов; обозначения те же, что в Таблице 3.

Отряд	G	St	G	St	G	St
	1 класс		2 класс		1 класс	
	$W_{(3,1)}$, CDS					
Ascomycetes	23	0,75	161	0,72	135	0,96
Basidiomycetes	8	0,75	21	0,72	21	0,96
Blastocladiomycota	0	0,75	2	0,72	2	0,96
Chytridiomycota	3	0,75	2	0,72	1	0,96
Zygomycota	4	0,75	5	0,72	5	0,96
	$W_{(3,1)}$, gene					
Ascomycetes	20	0,89	162	0,96	143	0,91
Basidiomycetes	7	0,89	21	0,96	22	0,91
Blastocladiomycota	0	0,89	2	0,96	2	0,91
Chytridiomycota	3	0,89	2	0,96	1	0,91
Zygomycota	3	0,89	5	0,96	5	0,91
	$W_{(3,3)}$, CDS					
Ascomycetes	4	0,82	156	0,97	169	0,60
Basidiomycetes	0	0,82	24	0,97	24	0,60
Blastocladiomycota	0	0,82	2	0,97	2	0,60
Chytridiomycota	1	0,82	4	0,97	4	0,60
Zygomycota	0	0,82	6	0,97	6	0,60
	$W_{(3,3)}$, gene					
Ascomycetes	160	0,72	30	0,89	157	0,97
Basidiomycetes	24	0,72	5	0,89	24	0,97
Blastocladiomycota	2	0,72	0	0,89	2	0,97
Chytridiomycota	5	0,72	1	0,89	4	0,97
Zygomycota	6	0,72	3	0,89	6	0,97

на 3 кластера и так далее. Таким образом, слоями графа являются наборы кластеров, полученные при делении точек на 2, 3, 4 и 5 кластеров.

Ребрами, представленными на графах стрелками, являются пути перераспределения генов при увеличении количества классов, на которые делится все множество генов, то есть при переходе к следующему слою графа. Жирными стрелками соединены классы, которые по большей части состо-

Таблица 5 — Распределение генов по четырем классам в разрезе отделов; обозначения те же, что в Таблице 3.

Отдел	1 класс		2 класс		3 класс		4 класс	
	G	St	G	St	G	St	G	St
	$W_{(3,1)}$, CDS							
Ascomycetes	144	0,98	161	0,93	144	0,90	16	0,80
Basidiomycetes	19	0,98	17	0,93	22	0,90	7	0,80
Blastocladiomycota	0	0,98	2	0,93	2	0,90	0	0,80
Chytridiomycota	1	0,98	2	0,93	1	0,9	3	0,80
Zygomycota	2	0,98	4	0,93	4	0,90	3	0,80
	$W_{(3,1)}$, gene							
Ascomycetes	162	0,94	73	0,87	81	0,90	14	0,85
Basidiomycetes	21	0,94	14	0,87	8	0,90	4	0,85
Blastocladiomycota	2	0,94	2	0,87	1	0,90	0	0,85
Chytridiomycota	2	0,94	0	0,87	1	0,90	3	0,85
Zygomycota	5	0,94	2	0,87	6	0,90	1	0,85
	$W_{(3,3)}$, CDS							
Ascomycetes	155	0,96	147	0,98	5	0,82	166	0,95
Basidiomycetes	24	0,96	24	0,98	0	0,82	23	0,95
Blastocladiomycota	2	0,96	2	0,98	0	0,82	2	0,95
Chytridiomycota	3	0,96	2	0,98	1	0,82	4	0,95
Zygomycota	6	0,96	5	0,98	0	0,82	6	0,95
	$W_{(3,3)}$, gene							
Ascomycetes	37	0,63	142	0,98	3	0,73	162	0,58
Basidiomycetes	5	0,63	22	0,98	1	0,73	22	0,58
Blastocladiomycota	0	0,63	2	0,98	0	0,73	2	0,58
Chytridiomycota	2	0,63	3	0,98	0	0,73	4	0,58
Zygomycota	3	0,63	2	0,98	0	0,73	6	0,58

ят из одних и тех же элементов.

На Рисунке 1 сверху представлен слоистый граф распределения генов семейства *atp* для словаря, построенного на последовательностях без интронов с шагом рамки считывания $t = 1$. При разделении множества точек методом динамических ядер на 2 класса основная часть точек, принадлежащих гену *atp*9 (189), оказалась в одном классе, а в другом классе

Таблица 6 — Распределение генов по четырем классам в разрезе отделов; обозначения те же, что в Таблице 3.

Отдел	1 класс		2 класс		3 класс		4 класс		5 класс	
	G	St	G	St	G	St	G	St	G	St
	$W_{(3,1)}$, CDS									
Ascomycetes	147	0,96	138	0,95	26	0,75	125	0,88	30	0,69
Basidiomycetes	16	0,96	16	0,95	4	0,75	20	0,88	8	0,69
Blastocladiomycota	2	0,96	0	0,95	0	0,75	2	0,88	1	0,69
Chytridiomycota	1	0,96	1	0,95	3	0,75	1	0,88	4	0,69
Zygomycota	1	0,96	0	0,95	1	0,75	4	0,88	5	0,69
	$W_{(3,1)}$, gene									
Ascomycetes	69	0,88	162	0,92	15	0,77	143	0,93	72	0,91
Basidiomycetes	12	0,88	19	0,92	6	0,77	19	0,93	10	0,91
Blastocladiomycota	2	0,88	2	0,92	0	0,77	0	0,93	0	0,91
Chytridiomycota	0	0,88	2	0,92	3	0,77	1	0,93	1	0,91
Zygomycota	2	0,88	5	0,92	2	0,77	3	0,93	3	0,91
	$W_{(3,3)}$, CDS									
Ascomycetes	116	0,96	164	0,95	5	0,77	155	0,92	31	0,82
Basidiomycetes	23	0,96	21	0,95	2	0,77	24	0,92	1	0,82
Blastocladiomycota	2	0,96	2	0,95	0	0,77	2	0,92	0	0,82
Chytridiomycota	0	0,96	4	0,95	1	0,77	3	0,92	3	0,82
Zygomycota	5	0,96	6	0,95	0	0,77	6	0,92	0	0,82
	$W_{(3,3)}$, gene									
Ascomycetes	142	0,97	3	0,8	27	0,84	165	0,3	2	0,78
Basidiomycetes	22	0,97	0	0,8	5	0,84	24	0,3	0	0,78
Blastocladiomycota	2	0,97	0	0,8	0	0,84	2	0,3	0	0,78
Chytridiomycota	2	0,97	1	0,8	0	0,84	4	0,3	0	0,78
Zygomycota	2	0,97	0	0,8	3	0,84	6	0,3	0	0,78

оказалась основная часть точек, принадлежащая генам *atp*6 (184) и *atp*8 (195). В скобках указано количество точек. Из этих двух вершин выходят стрелки, которые показывают в какие вершины переходят элементы при переходе к следующему слою. Также на графе присутствуют стрелки, выделенные жирным. Например при увеличении числа кластеров до $k = 3$, большая часть генов *atp*9, переходит в третью сверху вершину, а гены *atp*6

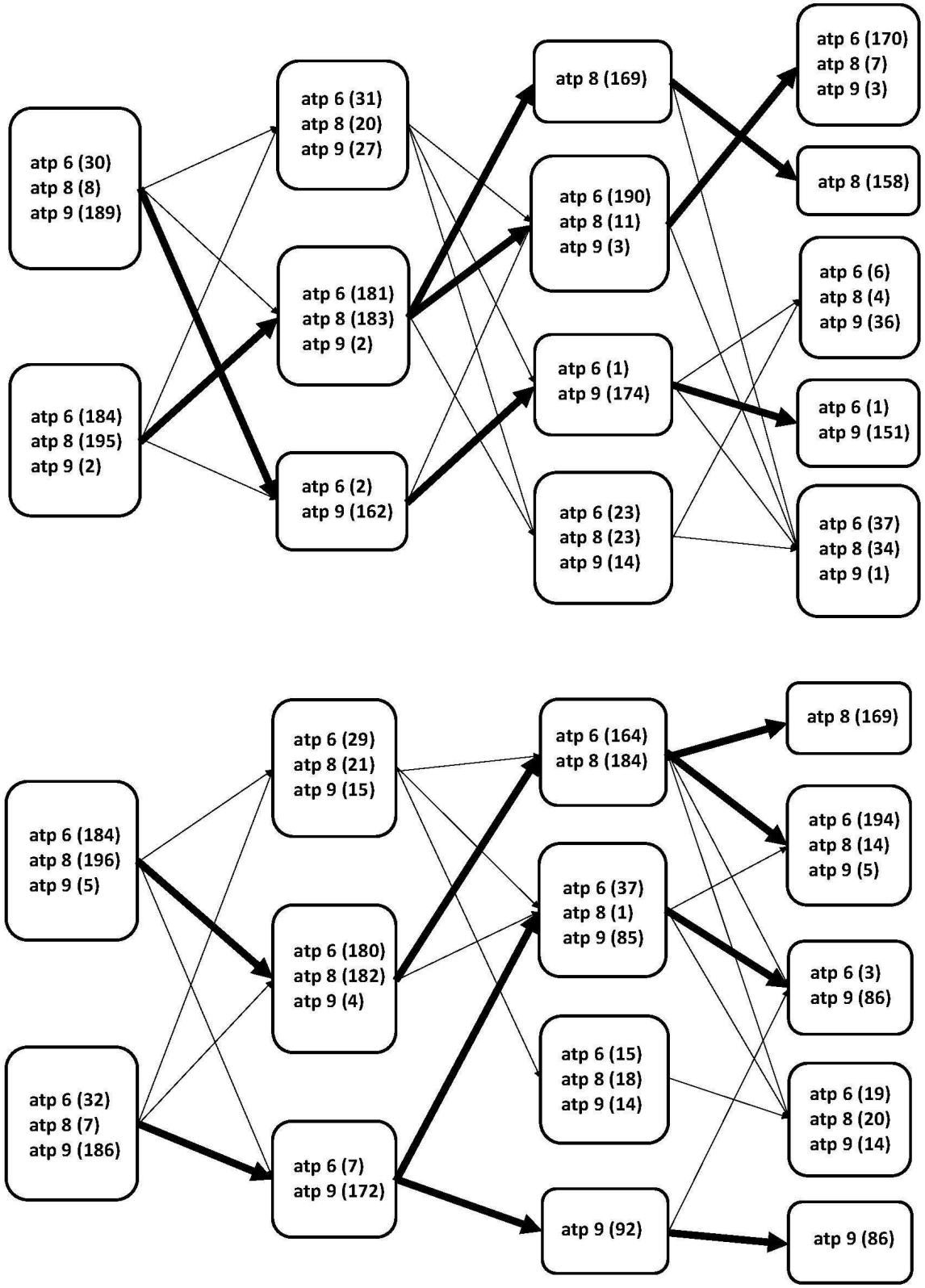


Рисунок 1 — Слоистый граф для частотных словарей CDS и Gene, $t = 1$.

и $atp8$ переходят во вторую вершину. Далее при переходе к $k = 4$ гены $atp9$ переходят в класс, в котором по большей части находятся только они.

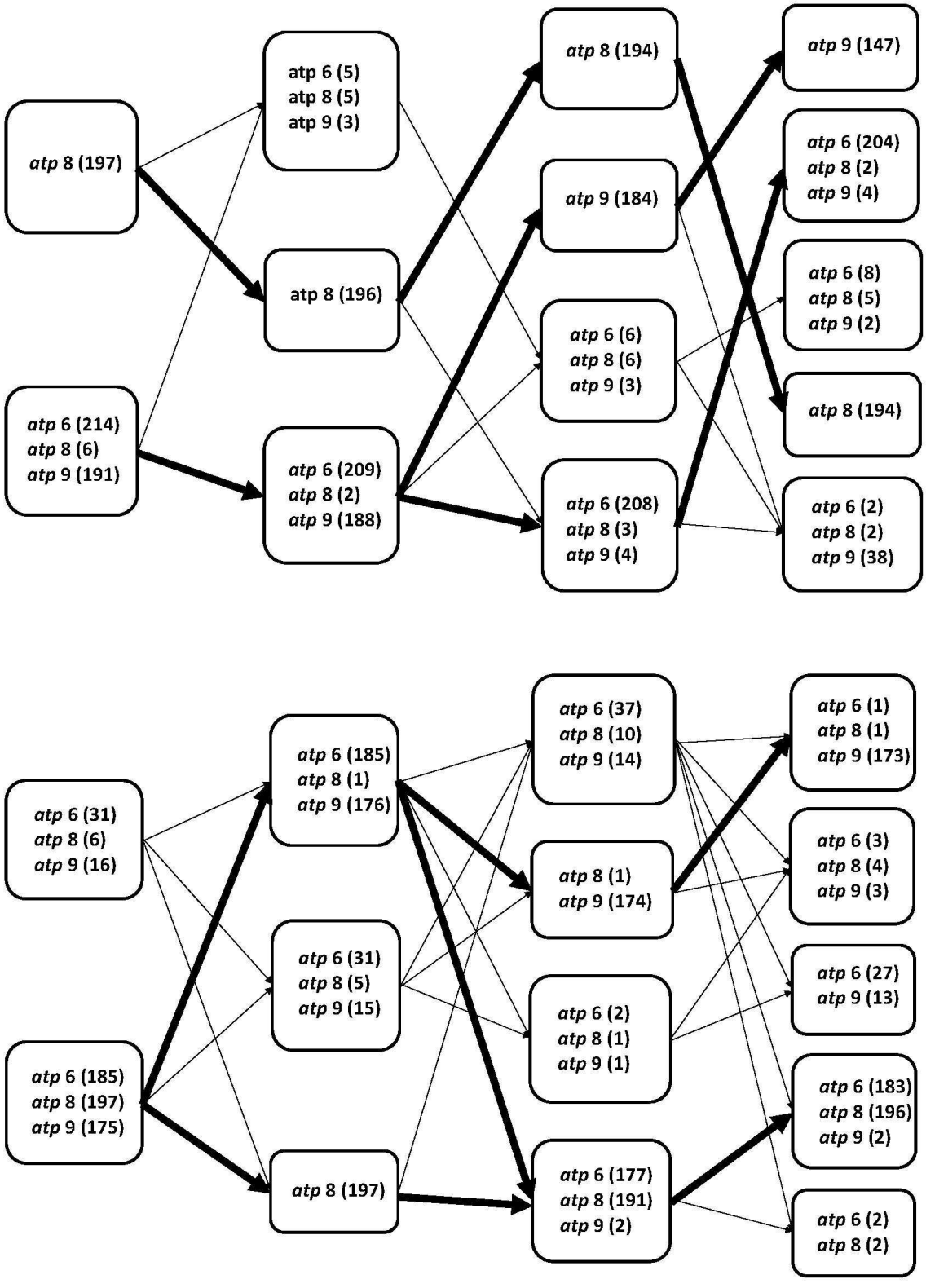


Рисунок 2 — Слоистый граф для частотных словарей CDS и Gene, $t = 3$.

Гены *atp*6 и *atp*8, находящиеся ранее в одном классе, теперь перешли в 2 разных класса. При переходе к $k = 5$, большая часть генов *atp*6, *atp*8 и *atp*9

попадают в разные классы. Далее для краткости будем говорить лишь о больших группах генов, собирающихся в кластер, то есть о тех группах куда попадает большинство точек.

На Рисунке 1 снизу представлен слоистый граф распределения генов семейства *atp* для словаря, построенного на последовательностях Gene с шагом рамки считывания $t = 1$. В этом случае гены *atp*6 и *atp*8 при $2 \leqslant k \leqslant 4$ делят один класс, в при $k = 5$ эти гены занимают отдельные классы. Гены *atp*9 при $k = 2$ и $k = 3$ занимает один класс, при $k = 4$ и $k = 5$ разделились почти пополам на 2 класса.

На Рисунке 2 сверху представлен слоистый граф распределения генов семейства *atp* для словаря, построенного на последовательностях CDS с шагом рамки считывания $t = 3$. Гены *atp*6 и *atp*9 занимают один класс при $k = 2$ и $k = 3$, далее при $k = 4$ и $k = 5$ эти гены разделяются и уже занимают разные классы. Гены *atp*8 при всех k занимают отдельный класс.

На Рисунке 2 снизу представлен слоистый граф распределения генов семейства *atp* для словаря, построенного на последовательностях с интронами и с экзонами с шагом рамки считывания $t = 3$. В этом случае при $k = 2$ гены *atp*6, *atp*8, *atp*9 занимают один класс, после чего разделяются. Далее при $k = 3$ гены *atp*6 и *atp*9 находятся в одном классе, а гены *atp*8 находятся в отдельном классе. При $k = 4$ уже гены *atp*6 и *atp*8 находятся в одном классе, а гены *atp*9 в отдельном. При $k = 5$ распределение аналогично тому, что и при $k = 4$.

Анализ распределения небольших групп генов выходит за рамки данной работы. Также в этой работе не представлен анализ видового состава больших групп генов. Интересно узнать одни и те же ли организмы находятся в этих группах, и как они переходят по классам при увеличении K.

3.3 Кластеризация словарей методом упругих карт

При проведении кластерного анализа были получены упругие карты распределения словарей в пространстве частот триплетов в так называемых внутренних координатах, полученных после выпрямления упругой

мембраны. Кластеры определялись по локальной плотности. Эта функция, описанная выше, показана в виде серого градиента горизонталей. Чем более плотным является распределение точек, тем более темным цветом обозначен участок карты.

На Рисунке 3 представлены полученные результаты кластеризации для всех исследуемых генов для словарей $W_{(3,1)}$ и $W_{(3,3)}$, построенных на CDS и gene последовательностях каждый. На этом рисунке цветом выделены изученные гены АТФ-синтаз: гены *atp*6 показаны синим цветом, гены *atp*8 — красным цветом, гены *atp*9 — желтым цветом. По этим рисункам хорошо видно, что вне зависимости от того, на каких словарях была построена карта, кластеры, выделяемые по локальной плотности, состоят из одинаковых генов.

На Рисунке 3(а) показано распределение генов для словарей $W_{(3,1)}$, построенных на последовательностях с интронами и экзонами. Очевидно, на этой упругой карте выделяются три хорошо различаемых кластера. При изучении функциональной принадлежности словарей, оказалось, что отдельные кластеры с большой точностью являются специфичными по составу: в левом кластере, выделенным желтым, находятся, как правило, гены *atp*9, в среднем, выделенным синим, — гены *atp*6, в правом, выделенным красным — гены *atp*8. Тем не менее, существует небольшое количество точек, нарушающее распределение, описанное выше.

Рисунок 3(б) показывает распределение генов для случая $W_{(3,3)}$, построенного на последовательностях Gene. Вид этой упругой карты отличается от вида упругой карты, представленной на Рисунке 3(а). Тем не менее, на этом рисунке также выделяются три хорошо различимых кластера. Каждый из кластеров оказался специфичным по составу: в кластере, находящемся в правом верхнем углу, находятся гены *atp*8, в среднем кластере — гены *atp*6, а в нижнем — гены *atp*9. Также существуют участки, которые нарушают данное распределение.

Рисунок 3(в), показывающий распределение генов для словарей $W_{(3,1)}$, построенных на последовательностях, не содержащих интронов, имеет по-

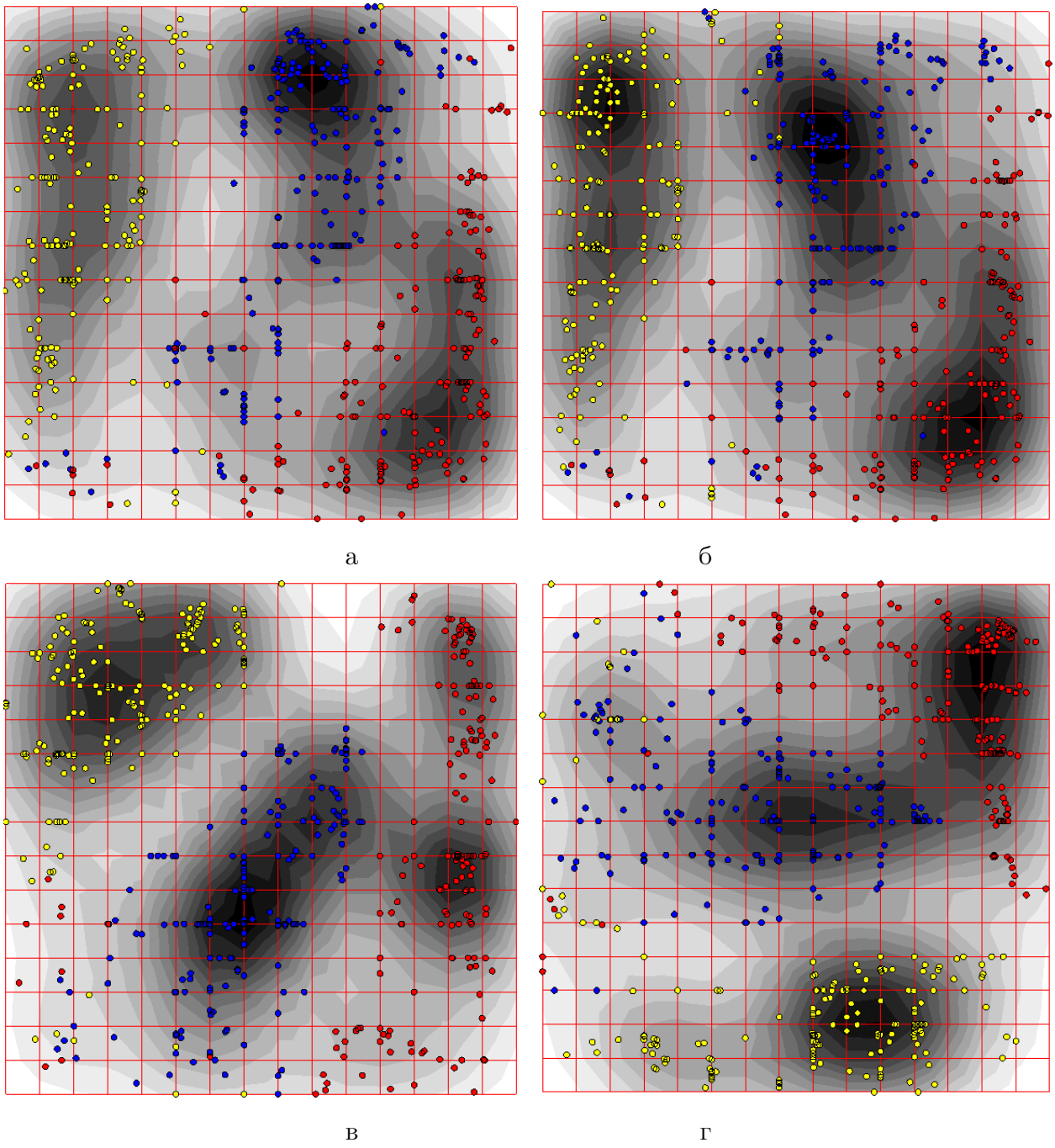


Рисунок 3 — Распределение генов в пространстве частот триплетов (синим обозначен ген *atp*6, красным — *atp*8, желтым — *atp*9): а — словарь $W_{(3,1)}$ Gene; б — словарь $W_{(3,3)}$ Gene; в — словарь $W_{(3,1)}$ CDS; г — словарь $W_{(3,3)}$ CDS.

хожий вид распределения с распределением генов, показанных на Рисунке 3(а). Заметим, что эти распределения показывают переход от последовательности CDS к последовательности Gene. В распределении генов, представленном на Рисунке 3(в) выделяются три хорошо различимых кластера. В кластере, находящемся слева собраны гены *atp*9, в среднем кластере — гены *atp*6, в кластере, находящемся справа — гены *atp*8. В таком распре-

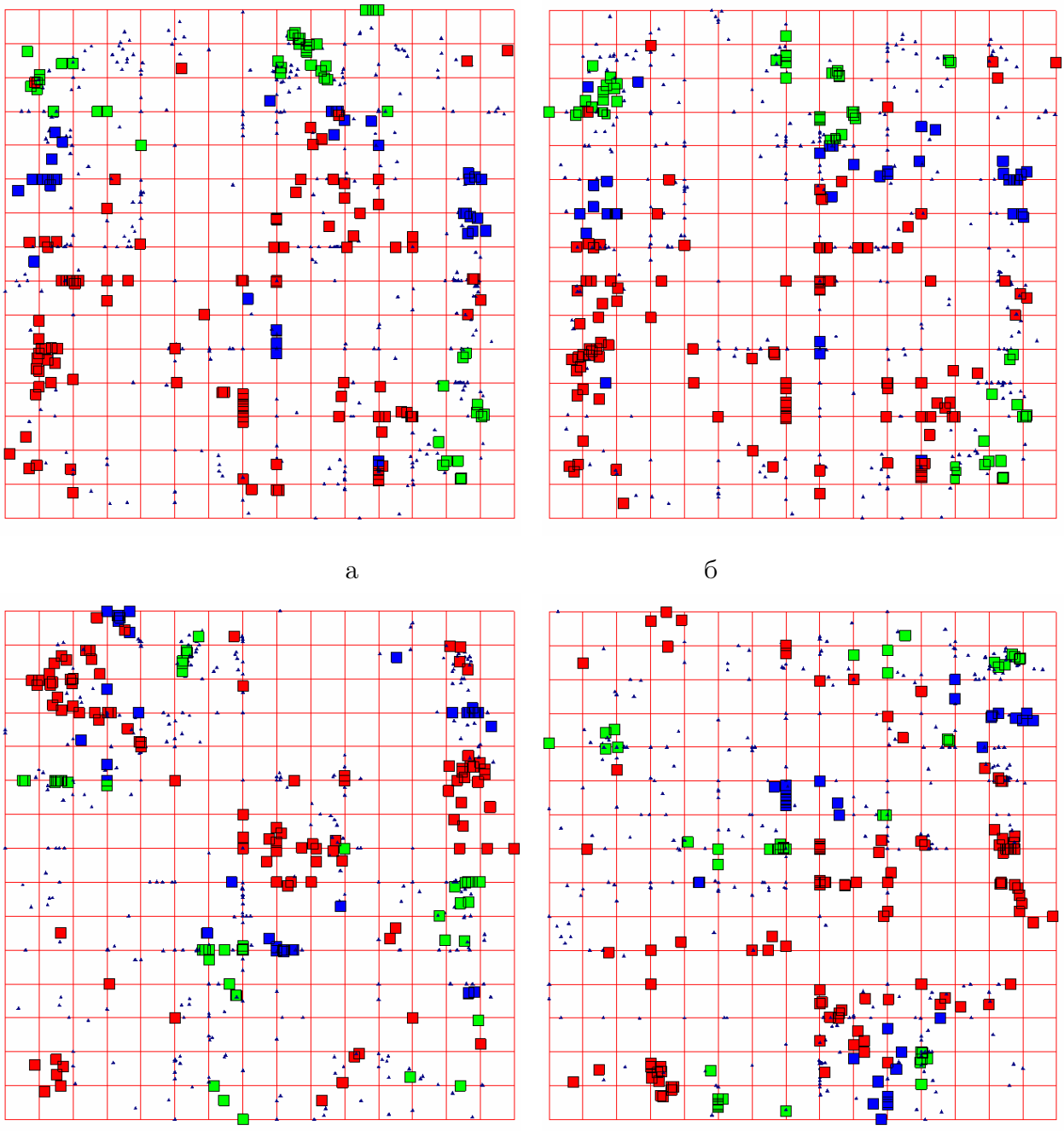
делении также существуют участки, нарушающие данное распределение.

На Рисунке 3(г) представлено распределение генов для словарей $W_{(3,3)}$, построенных на последовательности CDS. Хорошо видно, что на данном распределении выделяются три кластера, содержащие одинаковые гены. В кластере, находящемся в левом верхнем углу, собраны гены *atp*9, в среднем кластере — гены *atp*6, а в кластере справа — гены *atp*8. В этом распределении также, как и в остальных, существуют точки, нарушающие распределение генов, описанное выше. Кроме того, на Рисунке 3(г) можно заметить, что гены *atp*8 на самом деле образуют не один кластер, а два. Тем не менее, такое распределение генов *atp*8 может быть рассмотрено как один кластер, в зависимости от точности построения классификации.

Анализ точек, нарушающих распределение, описанное выше, выходит за рамки данной работы. Тем не менее, интересно проанализировать видовой состав этих групп точек и ответить на вопрос: одни и те же ли организмы являются исключениями из распределения генов?

Заметим, что переход от последовательностей, содержащих как экзоны, так и интроны (Gene), к последовательностям, из которых интроны были исключены (CDS), дает похожую картину кластеризации, однако существует отличие, заключающееся в более тонкой структуре кластеризации словарей на последовательностях, соответствующих зрелым РНК. На Рисунке 3(в) видно, что гены *atp*6 и *atp*8 на самом деле образуют два кластера, а не один. При этом в распределении генов *atp*9 нельзя надежно выделить кластеры. Тем не менее, вид распределения этих генов позволяет сделать предположение, что и для них наблюдается указанное разделение. На Рисунке 3(г) можно заметить, что гены *atp*8 явно разделены на два кластера, то же касается и *atp*6, но уже в меньшей степени, а гены *atp*9 нельзя надежно разделить на два кластера, что и наблюдалось в словаре $W_{(3,1)}$, построенном на той же последовательности.

В связи с этим, можно заключить, что различия словарей $W_{(3,1)}$ и $W_{(3,3)}$ весьма малы. Этот факт, возможно, означает, что различия статистических свойств кодирующих и некодирующих нуклеотидных последо-



а б в г

Рисунок 4 — Распределение генов в пространстве частот триплетов (синим обозначен подтип *Agaricomycotina*, красным — *Saccharomycotina*, зеленый — *Pezizomycotina*): а — словарь $W_{(3,1)}$ CDS; б — словарь $W_{(3,1)}$ Gene; в — словарь $W_{(3,3)}$ CDS; г — словарь $W_{(3,3)}$ Gene.

вательностей генома сильнее, чем различия между статистическими свойствами словарей, построенных с разным шагом перемещения рамки считывания.

3.4 Таксономический состав кластеров

Целью данной работы является выявление связи между структурой, функцией и таксономией. В предыдущем пункте мы рассматривали связь структуры и функции. В этом же пункте мы рассмотрим связь структуры и таксономии. Для этого просматривался таксономический состав кластеров: были выбраны три наиболее многочисленных подтипа, изученных организмов: *Agaricomucotina*, *Saccharomycotina* и *Pezizomycotina*, после чего было проверено распределение данных подтипов на упругой карте. На Рисунке 4 представлены результаты этого анализа.

Как видно из Рисунка 4, выбранные подтипы распределены по упругой карте достаточно равномерно. Ожидается, что остальные подтипы распределяются подобным образом, так как число некоторых видов, составляющих подтип, конечное и небольшое.

4 Обсуждение

В работе показано, что в результате нелинейной кластеризации частотных словарей методом упругих карт выделяются три кластера. Также исследовались признаки тех точек, которые образуют кластер; напомним, что кластеризация проводилась исключительно по частотам триплетов, но каждая точка представляет собой конкретный ген. Мы изучали функциональный состав кластеров, то есть отвечали на вопрос, верно ли, что в одном кластере собираются словари, соответствующие одинаковым генам. Кроме того, исследовался вопрос о таксономическом составе кластеров: верно ли, что в одном кластере собираются словари, соответствующие близким таксонам. Оказалось, что кластеры являются функционально специфичными, а связь с таксономией выражена крайне слабо.

Такое распределение словарей доказывает преобладание функции над таксономией. Тем не менее, это не окончательное доказательство преобладания, указанного выше. Для его получения следует ответить на следующие вопросы:

- Различимость классов;
- Использование не только Евклидовой метрики, но и других;
- Использование выборки организмов более однородной по отделам.

Подобный анализ, проведенный ранее по полным митохондриальным геномам животных, показал, наоборот, преобладание таксономии над функцией [5, 18] (также см. близкие, но выполненные в иной технике работы [19,20]). Подобные результаты были получены и при проведении анализа на хлоропластном генетическом материале.

4.1 Волатильные точки

При построении упругих карт словари, принадлежащие одному и тому же гену, как правило, образовывали кластер. Тем не менее, во всех вариантах частотных словарей, находились такие точки, которые при разделении на классы не оказывались в своем классе. Имеет смысл определить принадлежность таких словарей к таксономическим единицам. Кроме того,

возможно, стоит исключить из выборки такие словари и провести кластеризацию без них. Ожидается, что кластеризация таких данных будет более точной и устойчивой.

4.2 Влияние типа частотного словаря на кластеризацию генов

В данной работе была изучена связь между структурой (понимаемой как частотный словарь триплетов) нуклеотидной последовательности, функцией, кодируемой ей и таксономией ее носителей. В работе было рассмотрено четыре типа частотных словарей. Для каждого вида последовательностей — содержащим только экзоны и содержащим как экзоны, так и интроны — строились два различных вида частотных словарей: с шагом рамки считывания $t = 1$ (словарь $W_{(3,1)}$) и $t = 3$ (словарь $W_{(3,3)}$). Одной из задач данной работы является проверка того, насколько сильно влияние способа построения частотного словаря на кластеризацию генов.

Линейная кластеризация методом динамических ядер для последовательно возрастающего числа кластеров ($2 \leqslant K \leqslant 5$) позволила увидеть иерархический характер наблюдаемой кластеризации. Это означает, что по мере увеличения числа классов, в один класс попадают словари, все более тесно связанные. Было проверено по какому признаку словари образовывали все более тесные классы: по таксономическому или по функциональному. Проверка показала, что признаком близости объектов внутри класса являлась функциональная роль гена (см. Рисунки 1 – 2).

Наиболее заметное различие видно между кластеризацией диаметрально противоположный по способу построения словарей $W_{(3,1)}$ Gene и $W_{(3,3)}$ CDS: число ребер в первом случае намного больше, чем число ребер во втором случае (Рисунки 1 снизу и 2 сверху, соответственно). Также стоит отметить, что для словаря $W_{(3,3)}$ CDS характерно то, что большая часть точек, как правило, не смешивается друг с другом и не разделяется при переходе от слоя к слою. Обратная картина наблюдается в кластеризации словарей $W_{(3,1)}$ Gene: при $K = 4$ в одном классе смешались гены

*atp*6 и *atp*8, а ген *atp*9 разделился почти поровну на два класса; при $K = 5$ гены *atp*6 и *atp*8 находятся в разных классах, а большая часть точек, принадлежащих гену *atp*9 так и осталась разделенной почти пополам в двух классах. Кроме того, заметно, что число листьев (вершин из слоя с $K = 5$, имеющих только одно входящее ребро) заметно меньше для словаря $W_{(3,1)}$ Gene, чем для $W_{(3,3)}$ CDS. Впрочем, это справедливо в целом для перехода от частотных словарей, построенных на последовательностях Gene к словарям, построенным на последовательностях CDS.

Таксономический состав классов при использовании метода динамических ядер также был проверен. Влияния таксономии носителей генов на кластеризацию не было выявлено. В Таблицах 3 – 6 хорошо видно, что распределение таксонов по классам весьма близко к равномерному.

Кроме метода динамических ядер в работе была использована нелинейная кластеризация — метод упругих карт. На Рисунке 3 видно, что ни влияние интронов, ни влияние шага рамки считывания t, не является настолько сильным, чтобы исказить картину кластеризации. Тем не менее, кластеризация словарей, построенных на последовательностях CDS все же отличается от кластеризации словарей, построенных на последовательностях Gene. Для случая CDS-последовательностей наблюдается более детальная кластеризация: при кластеризации Gene-последовательностей четко выделяются 3 класса, тогда как при кластеризации CDS-последовательностей, выделенные классы разделяются на подклассы (см. Рисунок 3). Например, гены *atp*6 и *atp*8 на Рисунке 3(в) на самом деле можно разделить на 2 класса каждый, а гены *atp*8 на Рисунке 3(г) четко разделяются на 2 класса. Ответ на вопрос о причинах таких различий между данными генами выходит за рамки настоящей работы.

Для кластеризации, полученной методом упругих карт, также был проверен таксономический состав кластеров. Как и для кластеризации, полученной методом динамических ядер, таксоны были распределены на упругой карте весьма равномерно (см. Рисунок 4). Однако, следует учитывать тот факт, что выборка геномов была смещена по таксонам. Это

показано в Таблице 2: один из отделов представлен большим числом видов, один из них представлен почти впятеро меньшим числом видов, а 3 остальных отдела представлены очень малым числом видов.

4.3 Словарь $W_{(3,3)}$ CDS

Ранее был показан анализ частот триплетов для словаря $W_{(3,3)}$ CDS для генов семейства *atp*. Словарь, построенный на последовательностях, не имеющих интронов с шагом $t = 3$, можно рассматривать как словарь кодонов, так как последовательность без интронов эквивалентна зрелой РНК, готовой к трансляции, при этом рамка считывания двигается аналогично рибосоме.

Этот факт позволяет провести кластеризацию генов в пространстве с меньшей размерностью: частотный словарь кодонов можно легко преобразовать в частотный словарь аминокислот. Для этого следует объединить (сложить) частоты встречаемости в геноме синонимичных кодонов. Такая трансформация словарей может сократить размерность Евклидового пространства, в котором происходит их анализ, с 64 до 21, где координатами будут являться частоты аминокислот и стоп-кодона. Ожидается, что такой переход повысит точность классификации и ее устойчивость.

Кроме того, случай $t = 3$ порождает три разных словаря, которые отличаются стартовым положением рамки считывания. Эти словари также нуждаются в отдельном изучении. На таких отличиях в словарях строится методология *HMM*-анализа [21–24], а также выявления внутренней структурированности геномов [25–28].

4.4 Выбор генов

В данной работе изучалось распределение трех митохондриальных генов грибов семейства *atp*: *atp*6, *atp*8, *atp*9, отвечающих за окислительное фосфорилирование. Возможно, имеет смысл расширить набор генов, включенных в исследование. Таким образом, включение генов *nad*1-6, *nad*4L, *cox*1, *cox*2, *cox*3 и *cob* может принести новые знания о соотношении между

структурой, функцией и таксономией. Исследование распределения частотных словарей можно проводить

- на всех генах, включая гены группу *atp*;
- на группах генов по отдельности. Например, отдельно на группе *nad* и *cox*;
- отдельно на каждом из генов;
- на различных комбинациях отдельных генов или групп.

Также необходимо исследовать кластеризацию генов семейства *atp* отдельно. Иными словами, проверить структурированность, выделяемую на частотных словарях, построенных на одном гене (*atp*6, *atp*8 и *atp*9 по отдельности). В таком случае ожидается, что распределение точек и их кластеризация будет во многом похожа на кластеризацию, которая наблюдается на геномах органелл: ранее было установлено, что похожий анализ выявил превосходство таксономии носителей последовательности над функцией, кодируемой ей [6, 29, 30].

4.5 Выбор геномов

Рассуждения, приведенные в предыдущем пункте, также касаются выбора геномов при анализе соотношения в триаде *структура – функция – таксономия*. Нет никакой гарантии, что преобладание структуры над функцией наблюдается всегда, вне зависимости от выбора генома. Митохондриальные геномы являются хорошим объектом для такого рода исследований: они однородны по кодируемой в них функции, имеют одну хромосому. В любом случае, универсальность результатов данной работы должна быть проверена на других геномах, например, ядерных или хлоропластных.

5 Заключение

В ходе выполнения данной работы были выполнены все задачи, а поставленная цель достигнута:

1. Были построены частотные словари в четырех вариантах (на последовательностях CDS и Gene, с шагом $t = 1$ и $t = 3$ для каждой);
2. С помощью программы *ViDaExpert* были визуализированы данные, а также проведена кластеризация методом упругих карт, которая показала, что данные четко разбиваются на три кластера;
3. Была проведена кластеризация с помощью метода динамических ядер, посчитана устойчивость такой кластеризации и построены графы перехода точек между слоями при увеличении числа классов K.
4. Был проведен анализ состава каждого из классов: с точки зрения функции и таксономии и выявлены соответствующие характеристики распределений.

Анализ результатов данной работы показал, что в кластерах находятся словари, принадлежащие одному и тому же гену. Таким образом было доказано преобладание функции, кодируемой исследованными последовательностями, над таксономией их носителей. Нельзя сказать, что этот вывод является универсальным, то есть при проведении анализа последовательностей по частотному составу триплетов преобладание функции над таксономией может не выполняться на другом генетическом материале, например, на геномах хлоропластов или на геномах животных. Для того, чтобы утверждать об универсальности такого эффекта, необходимо провести дополнительные исследования.

СПИСОК СОКРАЩЕНИЙ

1. CDS (coding sequence) — кодирующая последовательность ДНК или РНК, соответствует последовательности аминокислот в белке;
2. ДНК — дезоксирибонуклеиновая кислота;
3. РНК — рибонуклеиновая кислота;
4. Gene — последовательность ДНК или РНК, включающая в себя как экзоны, так и интроны;
5. мРНК — матричная РНК;
6. иРНК — информационная РНК;
7. АТФ — аденозинтрифосфат;
8. АДФ — аденозиндифосфат;
9. ГТФ — гуанозинтрифосфат;
10. НАД — никотинамидадениндинуклеотид;
11. НАДФ — никотинамидадениндинуклеотидфосфат;
12. ФМН — флавинмононуклеотид;
13. ФАД — флавинадениндинуклеотид.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Основы молекулярной биологии клетки / Б Альбертс, Д Брей, К Хопкин et al. — М.: Бином, 2015.

[2] High variability of mitochondrial gene order among fungi / G. Aguileta, D. M. De Vienne, O. N. Ross et al. // *Genome biology and evolution.* — 2014. — Vol. 6, no. 2. — Pp. 451–465.

[3] *Smith, D. R.* The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs? / D. R. Smith // *Briefings in functional genomics.* — 2016. — Vol. 15, no. 1. — Pp. 47–54.

[4] *Deza, M. M.* Encyclopedia of distances / M. M. Deza, E. Deza // Encyclopedia of distances. — Springer, 2009. — Pp. 1–583.

[5] Genome structure of organelles strongly relates to taxonomy of bearers / M. Sadovsky, Yu. Putintseva, A. Chernyshova, V. Fedotova // International Conference on Bioinformatics and Biomedical Engineering / Springer. — 2015. — Pp. 481–490.

[6] *Sadovsky, M. G.* System Biology on Mitochondrion Genomes / M. G. Sadovsky, N. A. Zaitseva, Yu. A. Putintseva // The Third International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies. — 2011. — Pp. 61–66.

[7] *Gorban, A. N.* Classification of symbol sequences over their frequency dictionaries: towards the connection between structure and natural taxonomy / A. N. Gorban, T. G. Popova, M. G. Sadovsky // *Open Systems & Information Dynamics.* — 2000. — Vol. 7, no. 1. — Pp. 1–17.

[8] Information content of the frequency dictionaries, reconstruction, transformation and classification of dictionaries and genetic texts / A. N. Gorban, T. G. Popova, M. G. Sadovsky, D. C. Wunsch // Intelligent Engineering Systems Through Artificial Neural Networks. — American Society of Mechanical Engineers (ASME), 2001. — Pp. 657–663.

[9] *Бадмаев, Б. Б.* Сравнение таксономической структуры кормовых растений двух видов наземных беличьих Западного Забайкалья и степной флоры региона / Б. Б. Бадмаев // *Вестник Красноярского государственного аграрного университета*. — 2009. — no. 9.

[10] *Горбань, А. Н.* Классификация нуклеотидных последовательностей по частотным словарям обнаруживает связь между их структурой и таксономическим положением организмов / А. Н. Горбань, Т. Г. Попова, М. Г. Садовский // *Журнал общей биологии*. — 2003. — Vol. 64, no. 1. — Pp. 65–77.

[11] *Садовский, М.Г.* Выявление связи структуры и таксономии геномов хлоропластов методом динамических ядер / М.Г. Садовский, А.И. Чернышова // *Фундаментальные исследования*. — 2014. — Vol. 3, no. 11.

[12] *Сенашова, М.Ю.* Семикластерная структура геномов хлоропластов отражает филогению их носителей / М.Ю. Сенашова, М.Г. Садовский // *Международный журнал прикладных и фундаментальных исследований*. — 2016. — Vol. 12, no. 7. — Pp. 1167–1173.

[13] *Сенашова, М.Ю.* Пространственная структура геномов цианобактерий / М.Ю. Сенашова, М.Г. Садовский // *Международный журнал прикладных и фундаментальных исследований*. — 2017. — Vol. 11, no. 2. — Pp. 255–259.

[14] *Gorban, A. N.* Principal manifolds and graphs in practice: From molecular biology to dynamical systems / A. N. Gorban, A. Yu. Zinovyev // *International Journal of Neural Systems*. — 2010. — Vol. 20, no. 03. — Pp. 219–232. — PMID: 20556849. https://www.worldscientific.com/doi/abs/10.1142/S0129065710002383.

[15] *Gorban, A. N.* Principal Manifolds for Data Visualisation and Dimension Reduction / A. N. Gorban, A. Yu. Zinovyev // Lecture Notes in Computa-

tional Science and Engineering / Ed. by A N Gorban, B Kégl, D Wünsch, A Yu Zinovyev. — Berlin – Heidelberg – New York: Springer, 2007. — Vol. 58. — Pp. 153–176.

[16] *Gorban, A. N.* Fast and user-friendly non-linear principal manifold learning by method of elastic maps / A. N. Gorban, A. Yu. Zinovyev // 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, October 19-21, 2015. — 2015. — Pp. 1–9. https://doi.org/10.1109/DSAA.2015.7344818.

[17] *Fukunaga, K.* Introduction to statistical pattern recognition / K. Fukunaga. — London: Academic Press, 1990.

[18] *Sadovsky, M. G.* Chloroplasts and Cytoplasm: Structure and Functions / M. G. Sadovsky, M. Yu. Senashova, Yu. A. Putintseva // Chloroplasts and Cytoplasm: Structure and Functions. — Nova Science Publishers, Inc., 2018. — Pp. 25–95.

[19] A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes / Ch. Esser, N. Ahmadinejad, Ch. Wiegand et al. // *Molecular Biology and Evolution.* — 2004. — Vol. 21, no. 9. — Pp. 1643–1660.

[20] *Nadimi, M.* Mitochondrial comparative genomics and phylogenetic signal assessment of mtDNA among arbuscular mycorrhizal fungi / M. Nadimi, L. Daubois, M. Hijri // *Molecular phylogenetics and evolution.* — 2016. — Vol. 98. — Pp. 74–83.

[21] *De Fonzo, V.* Hidden Markov models in bioinformatics / V. De Fonzo, F. Aluffi-Pentini, V. Parisi // *Current Bioinformatics.* — 2007. — Vol. 2, no. 1. — Pp. 49–61.

[22] Bioinformatic Analysis Reveals High Diversity of Bacterial Genes for Laccase-Like Enzymes / L. Ausec, M. Zakrzewski, A. Goesmann et al. // *PLOS ONE.* — 2011. — 10. — Vol. 6, no. 10. — Pp. 1–9. https://doi.org/10.1371/journal.pone.0025724.

[23] *Zapatka, M.* Handbook of Statistical Bioinformatics. Series: Springer Handbooks of Computational Statistics. H. H.-S. Lu, B. Schölkopf, and H. Zhao (Editors) (2011). Berlin, Heidelberg: Springer-Verlag. 627 pages, 266.43, ISBN: 978-3-642-16344-9. / M. Zapatka // *Biometrical Journal.* — 2013. — 07. — Vol. 55.

[24] *Krogh, A.* A hidden Markov model that finds genes in *E.coli* DNA / A. Krogh, I. S. Mian, D. Haussler // *Nucleic Acids Research.* — 1994. — 11. — Vol. 22, no. 22. — Pp. 4768–4778. https://doi.org/10.1093/nar/22.22.4768.

[25] *Gorban, A. N.* Codon usage trajectories and 7-cluster structure of 143 complete bacterial genomic sequences / A. N. Gorban, T. G. Popova, A. Yu. Zinovyev // *Physica A: Statistical Mechanics and its Applications.* — 2005. — Vol. 353. — Pp. 365 – 387. http://www.sciencedirect.com/science/article/pii/S0378437105000828.

[26] *Gorban, A. N.* Seven clusters in genomic triplet distributions / A. N. Gorban, T. G. Popova, A. Yu. Zinovyev // *In Silico Biology.* — 2003. — Vol. 3, no. 4. — Pp. 471–482. http://content.iospress.com/articles/in-silico-biology/isb00110.

[27] *Sadovsky, M. G.* Eight clusters, synchrony of evolution and unique symmetry in chloroplast genomes: The offering from triplets / M. G. Sadovsky, M. Yu. Senashova, Yu. A. Putintseva // Chloroplasts and all-all-all. — Nova Publishers, Inc., 2018. — Pp. 49–178.

[28] *Садовский, М. Г.* Восьмикластерная структура геномов хлоропластов наземных растений / М. Г. Садовский, М. Ю. Сенашова, А. В. Малышев // *Журнал общей биологии.* — 2018. — Vol. 79, no. 2. — Pp. 124–134.

[29] Genome Structure of Organelles Strongly Relates to Taxonomy of Bearers / M. G. Sadovsky, Yu. A. Putintseva, A. A. Chernyshova, Vaselina Fedotova // Bioinformatics and Biomedical Engineering / Ed. by Francisco Ortuño, Ignacio Rojas. — Cham: Springer International Publishing, 2015. — Pp. 481–490.

[30] *De Novo* Assembly and Cluster Analysis of Siberian Larch Transcriptome and Genome / M. G. Sadovsky, Yu. A. Putintseva, V. V. Birukov et al. // Bioinformatics and Biomedical Engineering / Ed. by Françisco Ortuño, Ignaçio Rojas. — Cham: Springer International Publishing, 2016. — Pp. 455–464.

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ:
заведующий кафедрой

_____ В. А. Кратасюк
"22" июня 2020 г.

БАКАЛАВРСКАЯ РАБОТА

03.03.02 Физика

ВЫЯВЛЕНИЕ СВЯЗЕЙ МЕЖДУ ТАКСОНОМИЕЙ, ФУНКЦИЕЙ И ТРИПЛЕТНЫМ СОСТАВОМ МИТОХОНДРИАЛЬНЫХ ГЕНОВ НЕКОТОРЫХ ГРИБОВ

Руководитель: 16.06.2020 _____ д.ф.-м.н., проф. М. Г. Садовский
дата, подпись уч.степень, должность

Выпускник: 16.06.2020 _____ В. Д. Федотовская
дата, подпись

Красноярск 2020